Rates of Convergence in ERM

We've proved learnability for cpt hypo spaces. Next, we'll derive quantitative result on rate of convergence for $\mathcal{E}_{\text{sampl.}}$ which's crucial to prove uniform convergence / PAC-learnability & control capacity of hypo space.

(1) <u>For finite hypo space:</u>

<u>Lem.</u> For $\mathcal{H}$ finite, $\hat{\mu}_n$ λ-ERM learner for unbiased ERF $\hat{\mathcal{L}}_n$. If $Z_{v,n} := \lambda(\mu \| v) - (c_n \hat{\mathcal{L}}_n(v) + h_n(\mu))$ is subgaussian for $\forall$ $v \in \mathcal{H}$ with var. σ_n^2 (indept of $\mathcal{H}$).

Then: i) $\mathbb{E}(\sup_{\mathcal{H}} Z_{v,n}) \leq (2\sigma_n^2 \log|\mathcal{H}|)^{\frac{1}{2}}$.

ii) if $Z_{v,1}$ is sub-gaussian with var. $\sigma_1^2 = \sigma$. $\hat{\mathcal{L}}_n = \sum_1^n \ell(x_j|v)$. $c_n = \frac{1}{n}$ and $h_n(\mu) = h(\mu)$ in addition $\Rightarrow$

$\sigma_n^2 = \sigma^2/n$. $\mathbb{E}(\sup_{\mathcal{H}} Z_{v,n}) \leq (2\sigma^2 \log|\mathcal{H}|/n)^{\frac{1}{2}}$

iii) If $\mu \in \mathcal{H}$. $X_j \overset{i.i.d.}{\sim} \mu$. $\exists v_* \in \mathcal{H}$. $Z_{v^*,n} = 0$

with ii) $\Rightarrow \mathbb{E}(d(\mu \| \hat{\mu}_n)) \le 2\left(\frac{2\sigma^2 \log|\mathcal{H}|}{n}\right)^{\frac{1}{2}}$

Cor. $\mathbb{P}(d(\mu \| \hat{\mu}_n) \ge \varepsilon) \le 2\left(\frac{2\sigma^2 \log|\mathcal{H}|}{n\varepsilon^2}\right)^{\frac{1}{2}}$.

$\mathbb{P}(\mu \neq \hat{\mu}_n) \le 2(2\sigma^2 \log|\mathcal{H}|/n)^{\frac{1}{2}} / \inf_{\mathcal{H}/\{\mu\}} d(\mu \| v)$

Pf: By Markov inequi. and ($\mathcal{H}$ finite)

$\mathbb{P}(\mu \neq \hat{\mu}_n) = \mathbb{P}(d(\mu \| \hat{\mu}_n) \ge \inf_{\mathcal{H}/\{\mu\}} d(\mu \| v))$

Pf: i) LHS $= \mathbb{E}(\alpha^{-1} \log e^{\alpha \sup_{\mathcal{H}} Z_{v,n}})$

$\underset{\text{on } \log x}{\overset{\text{Jensen}}{\le}} \alpha^{-1} \log \mathbb{E}(e^{\alpha \sup_{\mathcal{H}} Z_{v,n}})$

$\le \alpha^{-1} \log \mathbb{E}(\sum_{\mathcal{H}} e^{\alpha Z_{v,n}})$

$\overset{\text{subgauss}}{\le} \alpha^{-1} \log \mathbb{E}(\sum_{\mathcal{H}} e^{\frac{1}{2}\sigma_n^2 \alpha^2})$

$= \alpha^{-1} \log(|\mathcal{H}| e^{\frac{1}{2}\sigma_n^2\alpha^2}) = \frac{\sigma_n^2 \alpha}{2} + \frac{\log|\mathcal{H}|}{\alpha}$

choose $\alpha = (2\log|\mathcal{H}|)^{\frac{1}{2}} / \sigma_n$ to optimize.

ii) Note only $\sum \ell(X_k | v)$ is random.

iii) From Rmk of error decompose:

$\mathbb{E}(d(\mu \| \hat{\mu}_n)) = 2\mathbb{E}(\sup_{\mathcal{H}} Z_{v,n}) + \varepsilon_{mod}(\mu)$.

Def: Given metric $\bar{L}$ on $\mathcal{H}$. $\bar{L} > 0$. Empirical process $\{Z_v\}_{v \in \mathcal{H}}$ is subgaussian process with

Var. $\bar{L}^2\lambda^{-2}$ if $\mathbb{E}(Z_v)=0$. $\forall v\in\mathcal{H}$ and

$Z_\mu - Z_v$ is subgaussian with Var. $\bar{L}^2\bar{\lambda}(\mu,v)^2$.

Rmk: i) It gives Z_μ kind of Lip. conti.

which's crucial when considering

cpt. hypo space (in balls)

ii) λ isn't necessary to equi. div. d

which measures success of learning

(some can't be metric. e.g. d_{KL}. even)

e.g. Note under cond. of Lem ii) above. We

have $Z_{v,n} = \frac{1}{n}\sum_1^n \ell(X_i|v) - \mathbb{E}_\mu^n \ell(X|v))$

We claim: $Z_v := Z_{v,1}$ is subgaussian for

$V\in\mathcal{H}=\{N(m,1)\}_{m\in[-1,1]}$. $\lambda(N(m,1),N(m',1))$

$=|m-m'|$. $X\sim N(m_0,1)$. Where $\ell(X|v)$

$=\log f_v(x) = \frac{1}{2}(x-m)^2+\log\sqrt{2\pi}$. $v=N(m,1)$.

Since $\mathbb{E}_X(\ell(X|N(m,1))) = \frac{1}{2}(m-m_0)^2+\frac{1}{2}+\log\sqrt{2\pi}$

$\Rightarrow Z_m = \frac{1}{2}(X-m)^2 - \frac{1}{2}(m-m_0)^2 - \frac{1}{2}$

So: $Z_m - Z_{m'} = 2(m'-m)X + \text{const}$. which

is Gaussian r.v.

Next, we want to control number of center of covering balls.

Def: For $\varepsilon > 0$. $\bar{d}$ metric on $\mathcal{U}$. ($N, D(\varepsilon) \triangleq N, D(\varepsilon, \bar{d}, \mathcal{U})$)

i) Covering number $N(\varepsilon)$ is smallest number of center v_j. s.t. $\mathcal{U} \subset \bigcup_1^{N(\varepsilon)} \bar{B}(v_j, \varepsilon)$

ii) Packing number $D(\varepsilon)$ is largest number of $v_j \in \mathcal{U}$. s.t. $\min_{i,j} \bar{d}(v_i, v_j) > \varepsilon$.

Lem: $D(2\varepsilon) \overset{i)}{\leq} N(\varepsilon) \overset{ii)}{\leq} D(\varepsilon)$.

Pf: i) For v_j. $j = 1, \dots, N(\varepsilon)$. $\bigcup_1^{N(\varepsilon)} \bar{B}(v_j, \varepsilon) \supset \mathcal{U}$.

And μ_j. $j = 1, \dots, D(2\varepsilon)$. packing pts.

At most one μ_j can be contained in some ε-ball $\bar{B}(v_i, \varepsilon)$ since its diameter is at most 2ε. So: $D(2\varepsilon) \leq N(\varepsilon)$

ii) $v_j \in \mathcal{U}$. $j = 1, \dots, D(\varepsilon)$. If $\{\bar{B}(v_j, \varepsilon)\}$ can't cover $\mathcal{U}$. i.e. $\exists\, v \notin \bigcup \bar{B}(v_j, \varepsilon)$.

We can set $v_{D(\varepsilon)+1} = v \Rightarrow$ Contradiction.

e.g. For $\Theta \subset \mathbb{R}^d$ cpt with $\Theta \subset B_r(\theta')$. We have $\mathcal{U} = \{\mu_\theta\}_{\theta \in \Theta}$ and $\bar{d} = d_\Theta$ satisfies:

$N(\varepsilon) \le D(\varepsilon) \le (1+\frac{2r}{\varepsilon})^d$.

Pf: For v_j. $j=1,\cdots D(\varepsilon)$. $\bar{B}(v_j, \frac{\varepsilon}{2})$ disjoint balls. s.t. $\cup \bar{B}(v_j, \frac{\varepsilon}{2}) \subset B_{r+\frac{\varepsilon}{2}}(\theta')$

$\Rightarrow D(\varepsilon) \cdot C_d (\frac{\varepsilon}{2})^d \le C_d (r+\frac{\varepsilon}{2})^d$

i.e. $D(\varepsilon) \le (1+2r/\varepsilon)^d$.

Thm. (Dudley's inequi on max entropy estimate)

$\{Z_v\}_{\mathcal{X}}$ is subgaussian with var. proxy $\bar{L}^2 \bar{d}^2$

If $(\mathcal{X}, \bar{d})$ is bdd. separable metric space. $v \mapsto Z_v(w)$ is conti. for $\mathbb{P}$-a.s. w. Then:

$\mathbb{E}(\sup_{\mathcal{X}} Z_v) \le 12 \sum_{k_0}^{\infty} 2^{-k} (\log(N(2^{-k}, \bar{L}\bar{d}, \mathcal{X})))^{\frac{1}{2}}$

where $k_0 \in \mathbb{Z}^+$ is largest s.t. $N(2^{-k_0}, \bar{L}\bar{d}, \mathcal{X}) = 1$.

Cor. $\mathbb{E}(\sup_{\mathcal{X}} Z_v) \le 24 \int_0^{\infty} (\log(N(\varepsilon, \bar{L}\bar{d}, \mathcal{X})))^{\frac{1}{2}} d\varepsilon$

$= 24 \bar{L} \int_0^{\infty} (\log(N(\varepsilon, \bar{d}, \mathcal{X})))^{\frac{1}{2}} d\varepsilon$

Pf: $12 \sum_{k_0}^{\infty} \square = 24 \sum_{k_0}^{\infty} \int_{2^{-(k+1)}}^{2^{-k}} (\log N(2^{-k}, \cdots))^{\frac{1}{2}}$

$\le 24 \sum_{k_0}^{\infty} \int_{2^{-(k+1)}}^{2^{-k}} (\log(N(\varepsilon, \bar{L}\bar{d}, \mathcal{X})))^{\frac{1}{2}} d\varepsilon$.

$\le 24 \int_0^{\infty} (\log(N(\varepsilon, \bar{L}\bar{d}, \mathcal{X})))^{\frac{1}{2}} d\varepsilon$.

And $N(\varepsilon, \bar{L}\bar{d}, \mathcal{X}) = N(\varepsilon/\bar{L}, \bar{d}, \mathcal{X})$. let $\varepsilon = \varepsilon/\bar{L}$. substitution.

Rmk: For ε large enough. $N(\varepsilon, \bar{\lambda}, \mathcal{U}) = 1$ So:

$\int_N^\infty \square = 0$ will not explode while $\int_0^\delta \square$ part is dangerous.

Pf: i) If $N(\varepsilon, \Sigma\bar{\lambda}, \mathcal{U}) = \infty$ for ε small. Then: It's nothing to prove. WLOG. $N(\varepsilon, \Sigma\bar{\lambda}, \mathcal{U}) < \infty$

By cont: $\sup_{\mathcal{U}} Z_v = \sup_{\mathcal{U}'} Z_v$ a.s. where $\mathcal{U}'$ is countable dense set of $\mathcal{U}$.

WLOG. let $\mathcal{U} = \{h_n\}_{n \in \mathbb{Z}^{\geq 0}}$ countable set.

2) Set $\mathcal{U}_s := \{h_n\}_{n \leq s}$. Choose $V_{k,j}$ of $\mathcal{U}_s$ $\Sigma\bar{\lambda}$-balls with $r = 2^{-k}$ covering $\mathcal{U}_s$.

let k_1 satisfy $\forall k \geq k_1$. $N(2^{-k}, \Sigma\bar{\lambda}, \mathcal{U}_s) = s$

then: $\{V_{k,j}\} = \{h_n\}_{n \leq s}$ in this case.

And k_0 satisfies $\forall k \leq k_0$, $N(2^{-k}, \Sigma\bar{\lambda}, \mathcal{U}_s) = 1$.

For $v \in \mathcal{U}_s$. let $Z_k(v) = V_{k,j}$ where $v \in \bar{B}^{\Sigma\bar{\lambda}}(V_{k,j}, 2^{-k}) \Rightarrow Z_{k_1}(v) = v$. $Z_{k_0}(v) = V_{k_0,1}$

So: $Z_v = \sum_{k_0}^{k_1} (Z_{Z_k(v)} - Z_{Z_{k-1}(v)}) + Z_{Z_{k_0}(v)}$ $= Z_{V_{k_0,1}}$

$\Rightarrow \mathbb{E}(\sup_{\mathcal{U}_s} Z_v) \leq \Sigma_{k_0}^{k_1} \mathbb{E}(\sup_{\mathcal{U}_s}(Z_{Z_k(v)} - Z_{Z_{k-1}(v)}))$

3) Over $\mathcal{U}_s$. they're at most M many

different $Z_{2^k(v)} - Z_{2_{k-1}(v)}$, where

$M = N(2^{-k}, \bar{L}\bar{\lambda}, \mathcal{H}_s) \cdot N(2^{-(k+1)}, \bar{L}\bar{\lambda}, \mathcal{H}_s)$

$\le D(\cdots, \mathcal{H}_s) D(\cdots, \mathcal{H}_s) \overset{\text{mono.}}{\le} D(\cdots, \mathcal{H}) D(\cdots, \mathcal{H})$

$\overset{\text{Lem.}^2}{\le} N(2^{-(k+1)}, \bar{L}\bar{\lambda}, \mathcal{H})^2$.

With $\bar{L}\bar{\lambda}(Z_k(v), Z_{k-1}(v))$

$\le \bar{L}(\bar{\lambda}(Z_k(v), v) + \bar{\lambda}(Z_{k-1}(v), v)) \overset{\text{def}}{\le} 3 \cdot 2^{-k}$

Apply Lem.$'$: $\mathbb{E}(\sup_{\mathcal{H}_s} Z) \le$

$\Sigma_{k_0}^{k_1} (2 (3 \cdot 2^{-k})^2 \log(N(2^{-(k+1)}, \bar{L}\bar{\lambda}, \mathcal{H})^2))^{\frac{1}{2}}$

Then use MCT. let $s \uparrow \infty$.

cor. Under condition of Thm above where $Z_{v,n}$

$= \lambda(\mu \| v) - (c_n \hat{\bar{L}}_n(v) + h_n(\mu))$ with $\hat{\bar{L}}_n(v)$

$= \sum_1^n \tilde{\ell}(X_j | v)$ unbiased. $c_n = \frac{1}{n}$. $X_k \overset{\text{i.i.d.}}{\sim} \mu$. $Z_v \overset{\Delta}{=} Z_{v,1}$

Then: for ERM-learner $\hat{\mu}_n$, if $\exists v^* \in \mathcal{H}$. s.t.

$Z_{v^*, n} = 0$. $\Rightarrow$ We have:

$\mathbb{E}(\lambda(\mu \| \hat{\mu}_n)) \le 48 \bar{L} n^{-\frac{1}{2}} \int_0^\infty (\log(N(\varepsilon, \bar{\lambda}, \mathcal{H})))^{\frac{1}{2}} d\varepsilon$

$+ \varepsilon_{n, \text{mod}}$

And for $\mu \in \mathcal{T} \subset \mathcal{H}$. $\Rightarrow \varepsilon_{n, \text{mod}} = 0$.

So: $\mathbb{P}(\lambda(\mu \| \hat{\mu}_n) > \varepsilon) \le \frac{48 \bar{L}}{\sqrt{n}\, \varepsilon} \int_0^\infty (\log(N(\varepsilon', \bar{\lambda}, \mathcal{H})))^{\frac{1}{2}} d\varepsilon'$.

Rmk: In prior example, we have: $N(\varepsilon) \leq (1 + 2r/\varepsilon)^d$. $\Rightarrow$ RHS is finite!

Besides, RHS doesn't depend on μ.

So for $\mathcal{J} = \mathcal{H}$, it's PAC-learnable w.r.t. $\ell_{\varepsilon\varepsilon}$ and my $\log \hat{L}_n$ here.

Pf: From Z_r to $Z_{n,r}$. We only need to replace L by $L/\sqrt{n}$. Apply Thm above.